Using the past to estimate sensory uncertainty

Ulrik Beierholm[1+*], Tim Rohe[2,3,4+], Ambra Ferrari[5], Oliver Stegle[6,7,8], Uta Noppeney[2,5,9]

[1]Psychology Department, Durham University, Durham, United Kingdom.

[2]Max Planck Institute for Biological Cybernetics, Tübingen, Germany

[3]Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany

[4]Department of Psychology, Friedrich-Alexander University Erlangen-Nuernberg, Erlangen, Germany

[5]Centre for Computational Neuroscience and Cognitive Robotics, University of Birmingham, Birmingham, United Kingdom.

[6] Max Planck Institute for Intelligent Systems, Tübingen, Germany

[7] European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany

[8] Division for Computational Genomics & Systems Genetics, German Cancer Research Center, Heidelberg, Germany

[9] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

[+] These authors contributed equally to this work.

[*]Correspondence should be addressed to ulrik.beierholm@durham.ac.uk

Abstract

To form a more reliable percept of the environment, the brain needs to estimate its own sensory uncertainty. Current theories of perceptual inference assume that the brain computes sensory uncertainty instantaneously and independently for each stimulus. We evaluated this assumption in four psychophysical experiments, in which human observers localized auditory signals that were presented synchronously with spatially disparate visual signals. Critically, the visual noise changed dynamically over time continuously or with intermittent jumps. Our results show that observers integrate audiovisual inputs weighted by sensory uncertainty estimates that combine information from past and current signals consistent with an optimal Bayesian learner that can be approximated by exponential discounting. Our results challenge leading models of perceptual inference where sensory uncertainty estimates depend only on the current stimulus. They demonstrate that the brain capitalizes on the temporal dynamics of the external world and estimates sensory uncertainty by combining past experiences with new incoming sensory signals.

Introduction

Perception has been described as a process of statistical inference based on noisy sensory inputs (Knill and Pouget, 2004, Knill and Richards, 1996). Key to this perceptual inference is the estimation and/or representation of sensory uncertainty. Most prominently, in multisensory perception a more reliable or 'Bayes-optimal' percept is obtained by integrating sensory signals that come from a common source weighted by their relative reliabilities (i.e., precision or inverse of variance) with less weight assigned to less reliable signals. Likewise, sensory uncertainty shapes observers' causal inference. It influences whether observers infer that signals come from a common cause and should hence be integrated or else be processed independently (Aller and Noppeney, 2019, Kording et al., 2007, Rohe et al., 2019, Rohe and Noppeney, 2015b, Rohe and Noppeney, 2015a, Rohe and Noppeney, 2016, Wozny et al., 2010, Acerbi et al., 2018). Indeed, accumulating evidence suggests that human observers are close to optimal in many perceptual tasks (though see (Acerbi et al., 2014, Drugowitsch et al., 2016, Shen and Ma, 2016, Meijer et al., 2019)) and weight signals approximately according to their sensory reliabilities (Alais and Burr, 2004, Ernst and Banks, 2002, Jacobs, 1999, Knill and Pouget, 2004, van Beers et al., 1999, Drugowitsch et al., 2014, Hou et al., 2019).

An unresolved question is how human observers compute their sensory uncertainty. Current theories and experimental approaches generally assume that observers access sensory uncertainty near-instantaneously and independently across briefly (≤200 ms) presented stimuli (Ma and Jazayeri, 2014, Zemel et al., 1998). At the neural level, theories of probabilistic population coding have suggested that sensory uncertainty may be represented instantaneously in the gain of the neuronal population response (Ma et al., 2006, Hou et al., 2019). Yet, in our natural environment, sensory noise often evolves at slow timescales. For instance, visual noise slowly varies when walking through a snow storm. Observers may capitalize on the temporal dynamics of the external world and use the past to inform current estimates of sensory uncertainty. In this alternative account, more reliable estimates of sensory uncertainty would be obtained by combining past estimates with current sensory inputs as predicted by Bayesian learning.

To arbitrate between these two critical hypotheses, we presented observers with audiovisual signals in synchrony but with a small spatial disparity in a sound localization task. Critically, the spatial standard deviation (STD) of the visual signal changed dynamically over time continuously (experiment 1-3) or discontinuously (i.e. with intermittent jumps; experiment 4). First, we investigated whether the influence of the visual signal location on observers' perceived sound location depended on the noise only of the current visual signal or also of past

visual signals. Second, using computational modeling and Bayesian model comparison, we formally assessed whether observers update their visual uncertainty estimates consistent with i. an instantaneous learner, ii. an optimal Bayesian learner or iii. an exponential learner.

Results

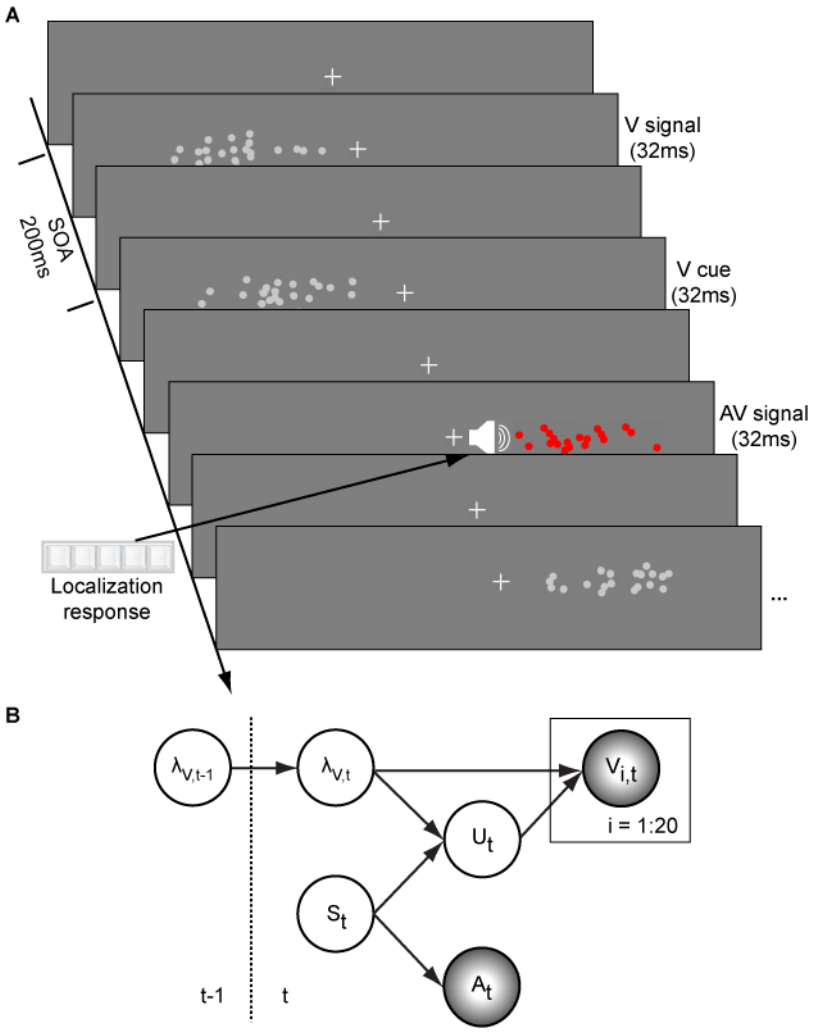

Figure 1. Audiovisual localization paradigm and Bayesian causal inference model for learning visual reliability. (**A**) Visual (V) signals (cloud of 20 bright dots) were presented every 200 ms for 32 ms. The cloud's location mean was temporally independently resampled from five possible locations (-10°, -5°, 0°, 5°, 10°) with an inter-trial asynchrony jittered between 1.4 and 2.8 s. In synchrony with the change in the cloud's mean location, the dots changed their colour and a sound was presented (AV signal) which the participants localized using five response buttons. The location of the sound was sampled from the two possible locations adjacent to the visual cloud's mean location (i.e. ± 5° AV spatial discrepancy). (**B**) The generative model for the Bayesian learner explicitly modelled the potential causal structures, i.e. whether visual (V_i) signals and an auditory (A) signal were generated by one common audiovisual source S_t , i.e. C = 1, or by two independent sources S_{Vt} and S_{At} , i.e. C = 2 (n.b. only the model component for the common source case is shown to illustrate the temporal updating, for complete generative model, see Figure 1-figure supplement 1). Importantly, the reliability (i.e., 1/variance) of the visual signal at time t (λ_t) depends on the reliability of the previous visual signal (λ_{t-1}) for both model components (i.e. common and independent sources).

In a spatial localization task, we presented participants with audiovisual signals in a series of four experiments, in which the physical visual noise changed dynamically over time either continuously or discontinuously (Figure 1). Visual (V) signals (clouds of 20 bright dots) were presented every 200 ms for a duration of 32 ms. The cloud's horizontal standard deviation (STD) varied over time at this temporal rate of 5 Hz either continuously (experiments 1-3) or

discontinuously with intermittent jumps (experiment 4). The cloud's location mean was temporally independently resampled from five possible locations (-10°, -5°, 0°, 5°, 10°) on each trial with the inter-trial asynchrony jittered between 1.4 and 2.8 s. In synchrony with the change in the cloud's mean location, the dots changed their colour and a sound was presented (AV signal). The location of the sound was sampled from the two possible locations adjacent to the visual cloud's mean location (i.e. ± 5° AV spatial disparity). Participants localized the sound and indicated their response using five response buttons.

The small audiovisual disparity enabled an influence of the visual signal location on the perceived sound location as a function of visual noise (Alais and Burr, 2004, Battaglia et al., 2003, Meijer et al., 2019). As a result, observers' visual uncertainty estimate could be quantified in terms of the relative weight of the auditory signal on the perceived sound location with a greater auditory weight indicating that observers estimated a greater visual uncertainty.

In the first three experiments, we used continuous sequences, where the visual cloud's STD changed periodically according to a sinusoid (n = 25; period = 30 s), a random walk (RW1; n = 33; period = 120 s) or a smoothed random walk (RW2; n = 19; period = 30 s; Figure 2). In an additional fourth experiment, we inserted abrupt increases or decreases into a sinusoidal evolution of the visual cloud's STD (n =18, period = 30 s, Figure 5). We will first describe the results for the three continuous sequences followed by the discontinuous sequence.

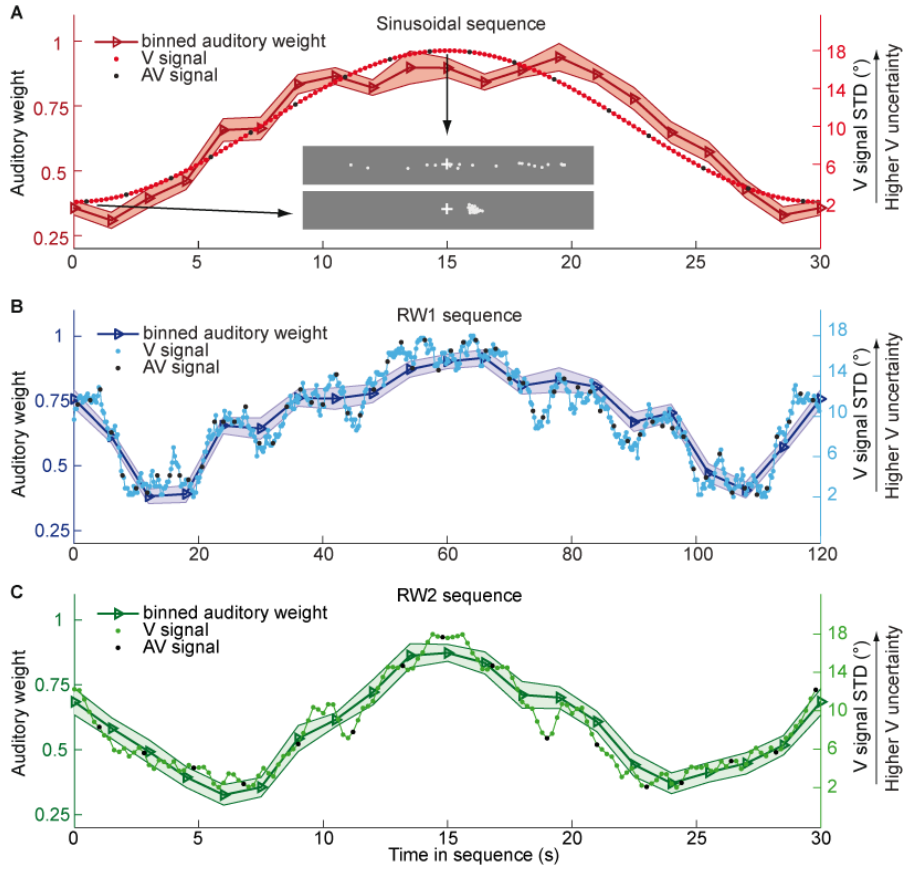

Figure 2. Time course of visual noise and relative auditory weights for continuous sequences of visual noise. The visual noise (i.e., STD of the cloud of dots, right ordinate) and the relative auditory weights (mean across participants ± SEM, left ordinate) are displayed as a function of time. The STD of the visual cloud was manipulated as (**A**) a sinusoidal (period 30s, N = 25), (**B**) a random walk (RW1, period 120s, N = 33) and (**C**) a smoothed random walk (RW2, period 30s, N = 19). The period for the RW1 sequence is 120 s, while the periods of the sinusoidal and RW2 is only 30 s. The overall dynamics as quantified by the power spectrum is faster for RW2 than RW1 (peak in frequency range [0 0.2] Hz: Sinusoid: 0.033 Hz, RW1: 0.025 Hz, RW2: 0.066 Hz). The RW1 and RW2 sequences were mirror-symmetric around the half-time (i.e., the second half was the reversed first half). The visual clouds were re-displayed every 200 ms (i.e., at 5 Hz). The trial onsets, i.e. audiovisual (AV) signals (color change with sound presentation, black dots), were interspersed with an inter-trial asynchrony jittered between 1.4 and 2.8 s. On each trial observers located the sound. The relative auditory weights were computed based on regression models for the sound localization responses separately for each of the 20 temporally adjacent bins that cover the entire period within each participant. The relative auditory weights vary between one (i.e. pure auditory influence on the localization responses) and zero (i.e. pure visual influence). For illustration purposes, the cloud of dots for the lowest (i.e., V signal STD = 2°) and the highest (i.e., V signal STD = 18°) visual variance are shown in (**A**).

We assigned the sound localization responses and the associated physical visual noise (i.e., the cloud's STD) to 20 (resp. 15 for experiment 4) temporally adjacent bins covering the entire period of each of the three sequences. Each experiment repeated the same 30 s (Sin, RW2) or 60s (RW1) period throughout the experiment resulting in ~ 32 periods for the RW2 and ~ 130 periods for the Sin and RW1 sequences. The trial and hence sound onsets were jittered with respect to this periodic evolution of the visual cloud's STD resulting in a greater effective sampling rate than expected for an inter-trial asynchrony of 1.4 – 2.8 s. In total, we assigned at least 44-87 trials to each bin (Supplementary file 1-Table 1). We quantified the auditory and visual influence on observers' perceived auditory location for each bin based on regression models (separately for each of the 20 temporally adjacent bins). For instance, for bin = 1 we computed:

$$(1)\ R_{A,trial,bin=1} = L_{A,trial,bin=1}\,ß_{A,bin=1} + L_{V,trial,bin=1}\,ß_{V,bin=1} + ß_{const,bin=1} + e_{trial,bin=1}$$

with $R_{A,trial,bin=1}$ = Localization response for trial t and bin 1; $L_{A,trial,bin=1}$ or $L_{V,trial,bin=1}$ = 'true' auditory or visual location for trial t and bin 1; $ß_{A,bin=1}$ or $ß_{V,bin=1}$ = auditory or visual weight for bin 1; $ß_{const,bin=1}$ = constant term; $e_{trial,bin=1}$ = error term. For each bin b, we thus obtained one auditory and one visual weight estimate. The relative auditory weight for a particular bin was computed as $w_{A,bin} = ß_{A,bin} / (ß_{A,bin} + ß_{V,bin})$.

Figure 2 and Figure 3 show the temporal evolution of the STD of the physical visual noise and observers' relative auditory weight indices $w_{A,bin}$. If observers estimate sensory uncertainty instantaneously, observer's relative auditory weight indices should closely track the visual cloud's STD (Figure 2). By contrast, we observed systematic biases: while the temporal

evolution of the physical visual noise was designed to be symmetrical for each time period, we observed a temporal asymmetry for w_A in all of the three experiments. For the monotonic sinusoidal sequence, w_A was smaller for the 1st half of each period, when visual noise increased, than the 2nd half, when visual noise decreased over time (Figure 3A). For the non-monotonic RW1 and RW2 sequences, we observed more complex temporal profiles, because the visual noise increased and decreased in each half. W_A was larger for increasing visual noise in the 1st as compared to the 2nd half ,while w_A was smaller for decreasing visual noise in the 1st as compared to the 2nd half (Figure 3B, C). These impressions were confirmed statistically in 2 (1st vs. flipped 2nd half) x 9 (bins) repeated measures ANOVAs (Table 1) showing a significant main effect of the 1st versus flipped 2nd half period for the sinusoidal ($F(1, 24) = 12.162$, $p = 0.002$, partial $\eta^2 = 0.336$) and the RW1 sequence ($F(1, 32) = 14.129$, $p < 0.001$, partial $\eta^2 = 0.306$). For the RW2 sequence, we observed a significant interaction ($F(4.6, 82.9) = 3.385$, $p = 0.010$, partial $\eta^2 = 0.158$), because the visual noise did not change monotonically within each half period. Instead, monotonic increases and decreases in visual noise alternated at nearly the double frequency in RW2 as compared to RW1. The asymmetry in the auditory weights' time course across the three experiments suggested that the visual noise in the past influenced observers' current visual uncertainty estimate resulting in smaller auditory weights for ascending visual noise and greater auditory weights for descending visual noise.

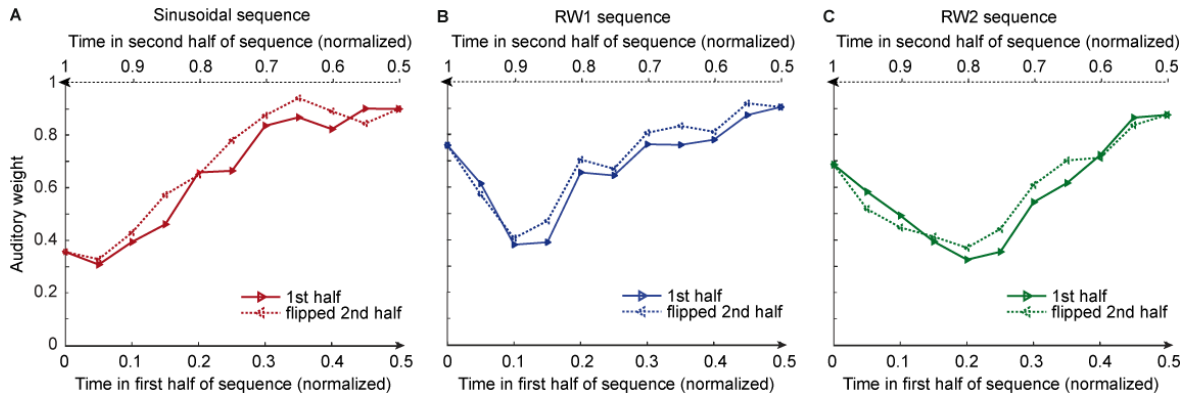

Figure 3. Observers' relative auditory weights for continuous sequences of visual noise. Relative auditory weights w_A of the 1st (solid) and the flipped 2nd half (dashed) of a period (binned into 20 bins) plotted as a function of the normalized time in the sinusoidal (red), the RW1 (blue) and the RW2 (green) sequences. Relative auditory weights were computed from auditory localization responses of human observers.

To further investigate the influence of past visual noise on observers' auditory weights, we estimated a regression model in which the relative auditory weights w_A for each of the 20 bins were predicted by the visual STD in the current bin and the difference in STD between the current and the previous bin (see equation (2)). Indeed, both the current visual STD ($p < 0.001$ for all three sequences; Sinusoid: $t(24)=15.767$, Cohen's $d=3.153$; RW1: $t(32)= 15.907$, Cohen's $d=2.769$; RW2: $t(18)=12.978$, Cohen's $d=2.977$, two sided one-sample t test against zero) and the difference in STD between the current and the previous bin (i.e. Sinusoid $t(24)= -3.687$, $p = 0.001$, Cohen's $d=-0.737$; RW1 $t(32)= -2.593$, $p = 0.014$, Cohen's $d=-0.451$; RW2 $t(18)=-2.395$, $p = 0.028$, Cohen's $d=-0.549$) significantly predicted observers' relative auditory weights (for complementary results of nested model comparison see Appendix 1 and Supplementary file 1-Table 5). Collectively, these results suggest that observers' visual

Table 1. Analyses of the temporal asymmetry of the relative auditory weights across the four sequences of visual noise using repeated measures ANOVAs with the factors sequence part (1st vs. flipped 2nd half), bin and jump position (only for the sinusoidal sequences with intermittent jumps).

	Effect	F	df1	df2	p	Partial η^2
Sinusoid	Part	12.162	1	24	0.002	0.336
	Bin	92.007	3.108	74.584	<0.001	0.793
	PartXBin	2.167	2.942	70.617	0.101	0.083
RW1	Part	14.129	1	32	0.001	0.306
	Bin	76.055	4.911	157.151	<0.001	0.704
	PartXBin	1.225	4.874	155.971	0.300	0.037
RW2	Part	2.884	1	18	0.107	0.138
	Bin	60.142	3.304	59.467	<0.001	0.770
	PartXBin	3.385	4.603	82.849	0.010	0.158
Sinusoid with intermittent jumps	Jump	28.306	2	34	<0.001	0.625
	Part	24.824	1	17	<0.001	0.594
	Bin	76.476	1.873	31.839	<0.001	0.818
	JumpXPart	0.300	2	34	0.743	0.017
	JumpXBin	8.383	3.309	56.247	<0.001	0.330
	PartXBin	1.641	3.248	55.222	0.187	0.088
	JumpXPartXBin	0.640	5.716	97.175	0.690	0.036

Note: The factor bin comprised 9 levels in the first three and 7 levels in the fourth sequence. In this sequence, the factor Jump comprised three levels. If Mauchly tests indicated significant deviations from sphericity ($p < 0.05$), we report Greenhouse-Geisser corrected degrees of freedom and p values.

uncertainty estimates (as indexed by the relative auditory weights w_A) depend not only on the current sensory signal, but also on the recent history of the sensory noise. These results were also validated in a control analysis that regressed out and thus accounted for potential influences of the previous visual location on observers' sound localization, suggesting that the effects of past visual uncertainty cannot be explained by effects of past visual location mean (Appendix 1, Figure 2-figure supplement 1, Supplementary file 1-Table 2-4).

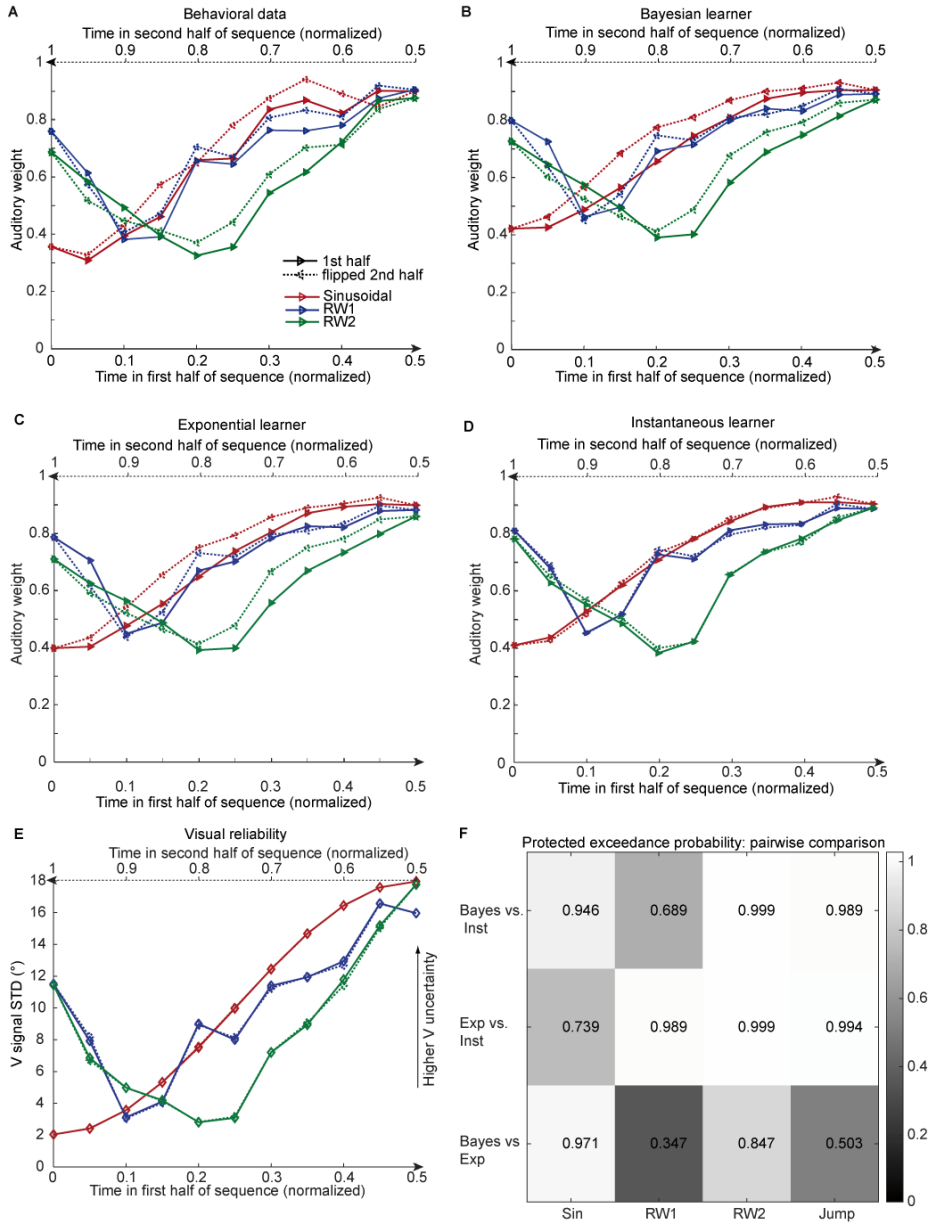

Figure 4. Observed and predicted relative auditory weights for continuous sequences of visual noise. Relative auditory weights w_A of the 1st (solid) and the flipped 2nd half (dashed) of a period (binned into 20 bins) plotted as a function of the normalized time in the sinusoidal (red), the RW1 (blue) and the RW2 (green) sequences. Relative auditory weights were computed from auditory localization responses of human observers **(A)**, Bayesian **(B)**, exponential **(C)** or instantaneous **(D)** learning models. For comparison, the standard deviation of the visual signal is shown in **(E)**. Please note that all models were fitted to observers' auditory localization responses (i.e. not the auditory weight w_A). (**F**) Bayesian model comparison – Random effects analysis: The matrix shows the protected exceedance probability (color coded and indicated by the numbers) for pairwise

comparisons of the Instantaneous (Inst), Bayesian (Bayes) and Exponential (Exp) learners separately for each of the four experiments. Across all experiments we observed that the Bayesian or the Exponential learner outperformed the Instantaneous learner (i.e. a protected exceedance probability > 0.94) indicating that observers used the past to estimate sensory uncertainty. However, it was not possible to arbitrate reliably between the Exponential and the Bayesian learner across all experiments (c.f. protected exceedance probability in bottom row).

To characterize how human observers use information from the past to estimate current sensory uncertainty, we compared three computational models that differed in how visual uncertainty is learnt over time (Figure 4): Model 1, the instantaneous learner, estimates visual uncertainty independently for each trial as assumed by current standard models. Model 2, the optimal Bayesian learner, estimates visual uncertainty by updating the prior uncertainty estimate obtained from past visual signals with the uncertainty estimate from the current signal. Model 3, the exponential learner, estimates visual uncertainty by exponentially discounting past uncertainty estimates. All three models account for observers' uncertainty about whether auditory and visual signals were generated by common or independent sources by explicitly modeling the two potential causal structures (Kording et al., 2007) underlying the audiovisual signals (n.b. only the model component pertaining to the 'common cause' case is shown in Figure 1B, for the full model see Figure 1-figure supplement 1). Models were fit individually to observers' data by sampling from the posterior over parameters for each observer (Table 2).

Table 2. Model parameters (median), absolute WAIC and relative ΔWAIC values for the three candidate models in the four sequences of visual noise.

Sequence	Model	σ_A	P_{common}	σ_0	κ or γ	WAIC	ΔWAIC
Sinusoid	Instantaneous learner	5.56	0.63	8.95	-	81931.2	109.9
	Bayesian learner	5.64	0.65	9.03	κ: 7.37	81821.3	0
	Exponential discounting	5.62	0.64	9.02	γ: 0.23	81866.9	45.6
RW1	Instantaneous learner	6.30	0.69	8.46	-	110051.2	89.0
	Bayesian learner	6.29	0.72	8.68	κ: 8.06	109962.2	0
	Exponential discounting	6.26	0.70	8.75	γ : 0.33	109929.9	-32.3
RW2	Instantaneous learner	6.36	0.72	10.79	-	62576.4	201.3
	Bayesian learner	6.49	0.78	10.9	κ: 6.7	62375.2	0
	Exponential discounting	6.46	0.73	11.0	γ : 0.25	62421.5	46.3
Sinusoid with intermittent jumps	Instantaneous learner	6.38	0.65	8.19	-	83891.4	94.9
	Bayesian learner	6.45	0.68	8.26	κ: 6.13	83796.5	0
	Exponential discounting	6.43	0.67	8.20	γ: 0.24	83798.1	1.64

Note: WAIC values were computed for each participant and summed across participants. A low WAIC indicates a better model. ΔWAIC is relative to the WAIC of the Bayesian learner.

We compared the three models in a fixed and random effects analysis (Penny et al., 2010, Rigoux et al., 2014) using the Watanabe-Akaike information criterion (WAIC) as appropriate for evaluating model samples (Gelman et al., 2014) (i.e., a low WAIC indicates a better model, a difference greater than 10 is considered very strong evidence for a model). In the fixed-effects analysis (see Table 2 for details), the Bayesian learner was substantially better than the instantaneous learner across all three experiments, but outperformed the exponential learner reliably only in the sinusoidal sequence. Likewise, the random-effects analysis based on hierarchical Bayesian model selection (Penny et al., 2010, Rigoux et al., 2014) showed a protected exceedance probability that was substantially greater for the Bayesian learner (Sin, RW2) or the exponential learner (RW1, RW2) than for the instantaneous learner (Figure 4F). However, the direct comparison between the Bayesian and the exponential learner did not provide consistent results across experiments. As shown in Figure 4 A and B, both the Bayesian and the exponential learner accurately reproduced the temporal asymmetry for the auditory weights across all three experiments.

From the optimal Bayesian learner we inferred observers' estimated rate of change in visual reliability (i.e. parameter $1/\kappa$). The sinusoidal sequence was estimated to change at a faster pace (median $\kappa = 7.4$ across observers, 95 percent confidence interval, 95%CI [4.8, 10.8] estimated via bootstrapping) than the RW1 sequence (median $\kappa = 8.1$, 95%CI [7.0,14.9]), but slower than the RW2 sequence (median $\kappa = 6.7$, 95%CI [4.4,11.2]) indicating that the Bayesian learner accurately inferred that visual reliability changed at different pace across the three continuous sequences (see legend of Figure 2). Likewise, the learning rates 1-γ of the exponential learner accurately reflect the different rates of change across the sequences (Sinusoid $\gamma = 0.23$, 95%CI [0.14, 0.28]; RW1: $\gamma = 0.33$, 95%CI [0.21, 0.38]; RW2: $\gamma = 0.25$, 95%CI [0.21, 0.29]). Both the Bayesian and the exponential learner thus estimated a smaller rate of change for the RW1 than for the sinusoidal sequence – though caution needs to be applied when interpreting these results given the extensive confidence intervals. Further, the learning rates of the exponential learner imply that observers gave the visual inputs presented 4.1 (Sinusoid), 5.4 (RW1) and 4.3 (RW2) seconds before the current stimulus 5% of the weight they assigned to the current visual input to estimate the visual reliability.

To further disambiguate between the Bayesian and the exponential learner, we designed a fourth experimental 'jump sequence' that introduced abrupt increases or decreases in physical visual noise at three positions into the sinusoidal sequence (Figure 5A). Using the same analysis

approach as for experiments 1-3, we replicated the temporal asymmetry for the auditory weights (Figure 5B). For all three 'jump positions' w_A was significantly smaller for the 1st half of each period, when visual noise increased, than the 2nd half, when visual noise decreased over time. The 3 (jump positions) x 2 (1st vs. flipped 2nd half) x 7 (bins) repeated measures ANOVA showed a significant main effect of 1st versus flipped 2nd period's half ($F(1,17) = 24.824$, $p < 0.001$, partial $\eta^2 = 0.594$), while this factor was not involved in any higher-order interaction (see Table 1). Further, in a regression model the current visual STD ($t(17) = 11.655$, $p < 0.001$, Cohen's $d = 2.747$) and the difference between current and previous STD ($t(17) = -4.768$, $p < 0.001$, Cohen's $d = -1.124$) significantly predicted the relative auditory weights. Thus, we replicated our finding that the visual noise in the past influenced observers' current visual uncertainty estimate as indexed by the relative auditory weights w_A.

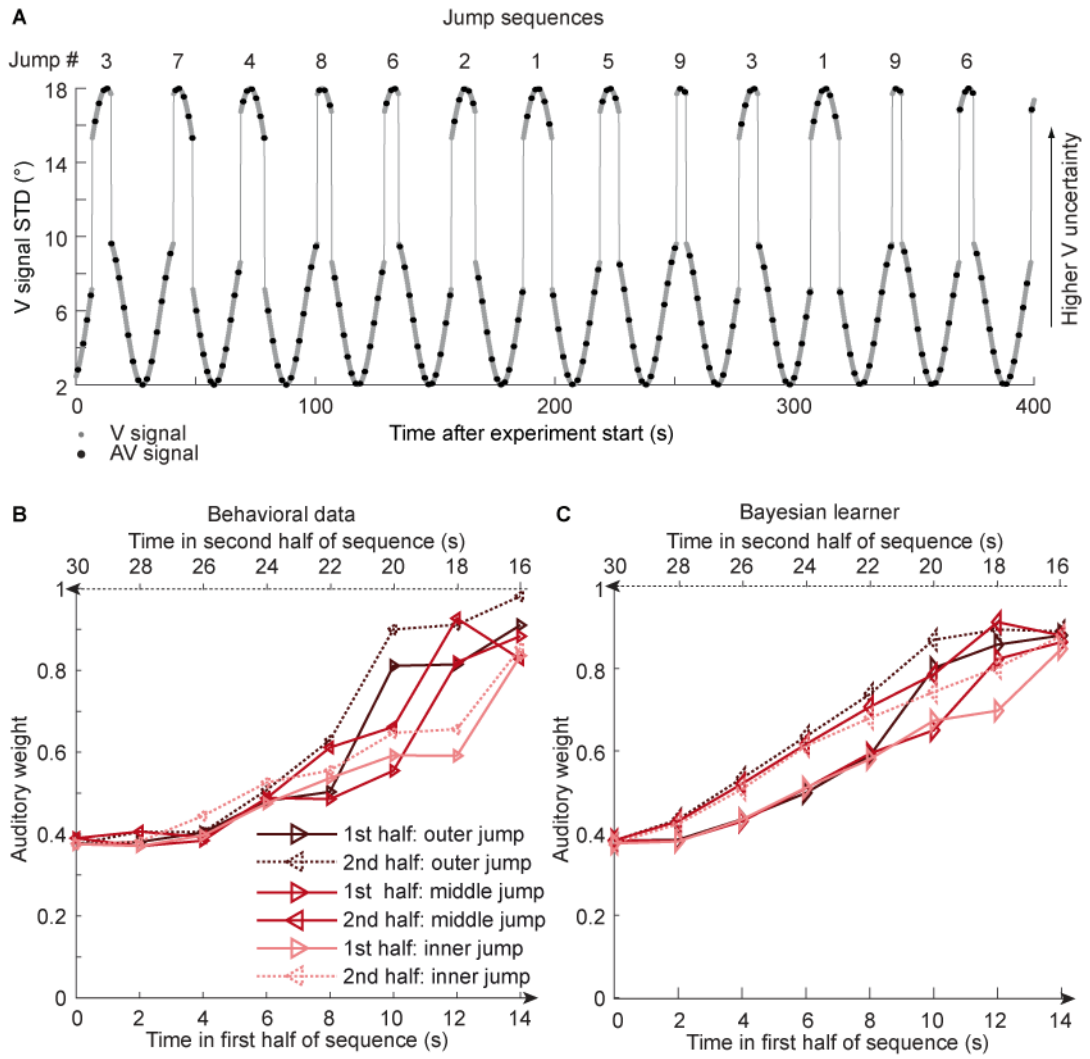

Figure 5. Time course of visual noise and relative auditory weights for sinusoidal sequence with intermittent jumps in visual noise (N = 18). (A) The visual noise (i.e., STD of the cloud of dots, right ordinate) is displayed as a function of time. Each cycle included one abrupt increase and decrease in visual noise. The sequence of visual clouds was presented every 200 ms (i.e., at 5 Hz) while audiovisual (AV) signals (black dots) were interspersed with an inter-trial asynchrony jittered between 1.4 and 2.8 s. **(B, C)** Relative auditory weights w_A of the 1st (solid) and the flipped 2nd half (dashed) of a period (binned into 15 bins) plotted as a function of the time in the sinusoidal sequence with intermitted inner (light gray), middle (gray) and outer (dark gray) jumps. Relative auditory weights were computed from auditory localization responses of human observers **(B)** and the Bayesian learning model **(C).** Please note that all models were fitted to observers' auditory localization responses (i.e. not the auditory weight w_A).

Bayesian model comparison using a fixed-effects analysis showed that both the Bayesian learner and the exponential learner substantially outperformed the instantaneous learner (see Table 2). However, consistent with our Bayesian model comparison results for the continuous sequences, the Bayesian learner did not provide a better explanation for observers' responses than the exponential learner (ΔWAIC = +2, see Table 2, Figure 5C and Figure 5-figure supplement 1A). Likewise, a random-effects analysis based on hierarchical Bayesian model selection showed that the Bayesian and the exponential learners outperformed the instantaneous learner, but again we were not able to adjudicate between the Bayesian and exponential learner (Figure 4F, see also methods and results in Appendix 1, Figure 5-figure supplement 2 and Supplementary file 1-Table 6 for further analyses justifying the choice of continuous learning models in the jump sequence).

In summary, across four experiments that used continuous and discontinuous sequences of visual noise, we have shown that the Bayesian or exponential learners outperform the instantaneous learner. However, across the four experiments we were not able to decide whether observers adapted to changes in visual noise according to a Bayesian or an exponential learner. The key feature that distinguishes between the Bayesian and the exponential learner is that only the Bayesian learner adapts dynamically based on its uncertainty about its visual reliability estimates. As a consequence, the Bayesian learner should adapt faster than the exponential learner to increases in physical visual noise (i.e. spread of the visual cloud) but slower to decreases in visual noise. From the Bayesian learner's perspective, the faster learning for increases in visual noise emerges because it is unlikely that visual dots form a large spread cloud under the assumption that the true visual spread of the cloud is small. Conversely, the Bayesian learner will adapt more slowly to decreases in visual variance, because under the assumption of a visual cloud with a large spread visual dots may form a small cloud by chance. Indeed, previous research has shown that observers adapt their variance estimates faster for changes from small to large than for changes from large to small variance (Berniker et al., 2010). However, these results have been shown for learning about a hidden variable such as the

prior that defines the spatial distribution from which an object's location is sampled. In our study, we manipulated the variance of the likelihood, i.e. the variance of the clouds of dots.

Asymmetric differences in adaptation rate between the exponential and the Bayesian learner should thus be amplified if we increase observer's uncertainty about its visual reliability estimate by reducing the number of dots of the visual cloud from 20 to 5 dots. Based on simulations, we therefore explored whether we could experimentally discriminate between the Bayesian and exponential learner using continuous sinusoidal or discontinuous 'jump' sequences with visual clouds of only 5 dots. For the two sequences, we simulated the sound localization responses of 12 observers based on the Bayesian learner model and fitted the Bayesian and exponential learner models to the responses of each simulated Bayesian observer. Figure 6 shows observers' auditory weights indexing their estimated visual reliability across time that we obtained from the fitted responses of the Bayesian (blue) and the exponential learner (green). The simulations reveal the characteristic differences in how the Bayesian and the exponential learner adapt their visual uncertainty estimates to increases and decreases in visual noise. As expected, the Bayesian learner adapts its visual uncertainty estimates faster than the exponential learner to increases in visual noise, but slower to decreases in visual noise. Nevertheless, these differences are relatively small, so that the difference in mean log likelihood between the Bayesian and exponential learner is only -1.82 for the sinusoidal sequence and -2.74 for the jump sequence.

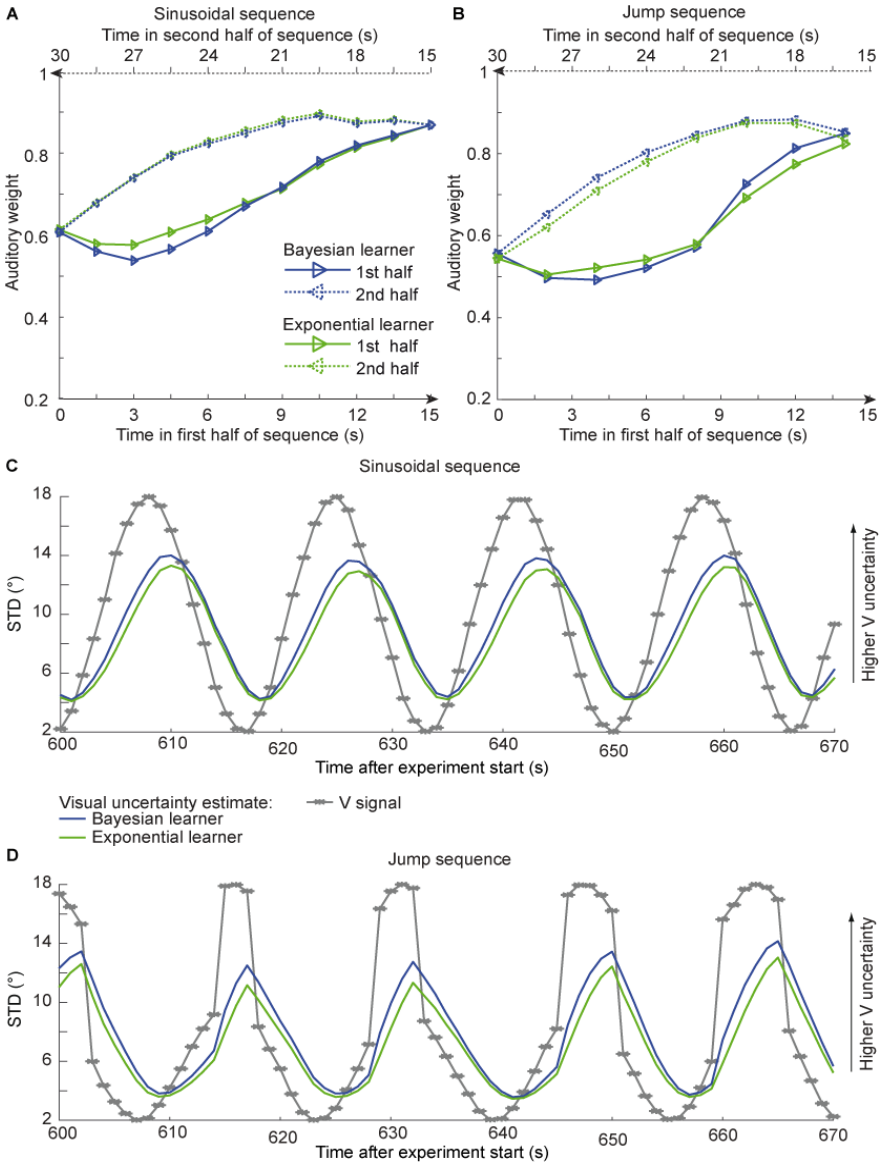

Figure 6. Time course of the relative auditory weights, the standard deviation of the visual cloud and the standard deviation of the visual uncertainty estimates. (A) Relative auditory weights w_A of the 1st (solid) and the flipped 2nd half (dashed) of a period (binned into 15 bins) plotted as a function of the time in the sinusoidal sequence. Relative auditory weights were computed from the predicted auditory localization responses of the Bayesian (blue) or exponential (green) learning models fitted to the simulated localization responses of a Bayesian learner based on visual clouds of 5 dots. **(B)** Relative auditory weights w_A computed as in (A) for the sinusoidal sequence with intermitted jumps. Only the outer-most jump (black in Figure 5B/C and Figure 5-figure supplement 1) is shown. **(C, D)** Standard deviation (STD) of the visual cloud of 5 dots (grey) and the STD of observers' visual uncertainty as estimated by the Bayesian (blue) and exponential (green) learners (that were fitted to the simulated localization responses of a Bayesian learner) as a function of time for the sinusoidal sequence (C) and in the sinusoidal sequence with intermitted jumps (D). Note that only an exemplary time course from 600-670 s after the experiment start is shown.

Next, we investigated whether our experiments successfully mimicked situations in which observers benefit from integrating past and current information to estimate their sensory uncertainty. We compared the accuracy of the instantaneous, exponential and Bayesian learner's visual uncertainty estimates in terms of their mean absolute deviation (in percentage)

from the true variance. For Gaussian clouds of 20 dots, the instantaneous learner's error in the visual uncertainty estimates of 21.7 % is reduced to 13.7 % and 14.9% for the exponential and Bayesian learners, respectively (with best fitted $\gamma = 0.6$, in the sinusoidal sequence). For Gaussian clouds composed of only 5 dots, the exponential and Bayesian learners even cut down the error by half (i.e. 46.8 % instantaneous learner, 29.5 % exponential learner, 23.9 % Bayesian learner, with best fitted $\gamma = 0.7$).

Collectively, these simulation results suggest that even in situations in which observers benefit from combining past with current sensory inputs to obtain more precise uncertainty estimates, the exponential learner is a good approximation of the Bayesian learner, making it challenging to dissociate the two experimentally based on noisy human behavioural responses.

Discussion

The results from our four experiments challenge classical models of perceptual inference where a perceptual interpretation is obtained using a likelihood that depends solely on the current sensory inputs (Ernst and Banks, 2002). These models implicitly assume that sensory uncertainty (i.e., likelihood variance) is instantaneously and independently accessed from the sensory signals on each trial based on initial calibration of the nervous system (Jacobs and Fine, 1999). Most prominently, in the field of cue combination it is generally assumed that sensory signals are weighted by their uncertainties that are estimated only from the current sensory signals (Alais and Burr, 2004, Ernst and Banks, 2002, Jacobs, 1999) (but see (Mikula et al., 2018, Triesch et al., 2002)).

By contrast, our results demonstrate that human observers integrate inputs weighted by uncertainties that are estimated jointly from past and current sensory signals. Across the three continuous and the one discontinuous jump sequences, observers' current visual reliability estimates were influenced by visual inputs that were presented 4-5 s in the past albeit their influence amounted to only 5% of the current visual signals.

Critically, observers adapted their visual uncertainty estimates flexibly according to the rate of change in the visual noise across the experiments. As predicted by both Bayesian and exponential learning models, observers' visual reliability estimates relied more strongly on past sensory inputs, when the visual noise changed more slowly across time. While observers did not explicitly notice that each of the four experiments was composed of repetitions of temporally symmetric sequence components, we cannot fully exclude that observers may have implicitly learnt this underlying temporal structure. However, implicit or explicit knowledge of this repetitive sequence structure should have given observers the ability to predict and preempt future changes in visual reliability and therefore attenuated the temporal lag of the visual reliability estimates. Put differently, our experimental choice of repeating the same sequence component over and over again in the experiment cannot explain the influence of past signals on observers' current reliability estimate, but should have reduced or even abolished it.

Importantly, the key feature that distinguishes the Bayesian from the exponential learner is how the two learners adapt to increases versus decreases in visual noise. Only the Bayesian learner represents and accounts for its uncertainty about its visual reliability estimates. As compared to the exponential learner, it should therefore adapt faster to increases but slower to decreases in visual noise (e.g. see Berniker et al. (2010)). Our simulation results show this profile qualitatively, when the learner's uncertainty about its visual reliability estimate is increased by reducing the number of dots (see Figure 6). But even for visual clouds of five dots, the

differences in learning curves between the Bayesian and exponential learner are very small making it difficult to adjudicate between them given noisy observations from real observers. Unsurprisingly, therefore, Bayesian model comparison showed consistently across all four experiments that observers' localization responses can be explained equally well by an optimal Bayesian and an exponential learner. These results converge with a recent study showing that learning about a hidden variable such as observers' priors can be accounted for by an exponential averaging model (Norton et al., 2019).

Collectively, our experimental and simulation results suggest that under circumstances where observers substantially benefit from combining past and current sensory inputs for estimating sensory uncertainty, optimal Bayesian learning can be approximated well by more simple heuristic strategies of exponential discounting that update sensory weights with a fixed learning rate irrespective of observers' uncertainty about their visual reliability estimate (Ma and Jazayeri, 2014, Shen and Ma, 2016). Future research will need to assess whether observers adapt their visual uncertainty estimates similarly if visual noise is manipulated via other methods such as stimulus luminance, duration or blur.

From the perspective of neural coding, our findings suggest that current theories of probabilistic population coding (Beck et al., 2008, Ma et al., 2006, Hou et al., 2019) may need to be extended to accommodate additional influences of past experiences on neural representations of sensory uncertainties. Alternatively, the brain may compute sensory uncertainty using strategies of temporal sampling (Fiser et al., 2010).

In conclusion, our study demonstrates that human observers do not access sensory uncertainty instantaneously from the current sensory signals alone, but learn sensory uncertainty over time by combining past experiences and current sensory inputs as predicted by an optimal Bayesian learner or approximate strategies of exponential discounting. This influence of past signals on current sensory uncertainty estimates is likely to affect learning not only at slower timescales across trials (i.e. as shown in this study), but also at faster timescales of evidence accumulation within a trial (Drugowitsch et al., 2014). While our research unravels the impact of prior sensory inputs on uncertainty estimation in a cue combination context, we expect that they reveal fundamental principles of how the human brain computes and encodes sensory uncertainty.

Methods

Participants

76 healthy volunteers participated in the study after giving written informed consent (40 female, mean age 25.3 years, range 18-52 years). All participants were naïve to the purpose of the study. All participants had normal or corrected-to normal vision and reported normal hearing. The study was approved by the human research review committee of the University of Tuebingen (approval number 432 2007 BO1) and the research review committee of the University of Birmingham (approval number ERN_11-0470P).

Stimuli

The visual spatial stimulus was a Gaussian cloud of twenty bright grey dots (0.56° diameter, vertical standard deviation 1.5°, luminance 106 cd/m^2) presented on a dark grey background (luminance 62 cd/m^2, i.e. 71% contrast). The auditory spatial cue was a burst of white noise with a 5ms on/off ramp. To create a virtual auditory spatial cue, the noise was convolved with spatially specific head-related transfer functions (HRTFs). The HRTFs were pseudo-individualized by matching participants' head width, heights, depth and circumference to the anthropometry of subjects in the CIPIC database (Algazi et al., 2001). HRTFs from the available locations in the database were interpolated to the desired locations of the auditory cue.

Experimental design and procedure

In a spatial ventriloquist paradigm, participants were presented with audiovisual spatial signals. Participants indicated the location of the sound by pressing one of 5 spatially corresponding buttons and were instructed to ignore the visual signal. Participants did not receive any feedback on their localization response. The visual signal was a cloud of 20 dots sampled from a Gaussian. The visual clouds were re-displayed with variable horizontal standard deviations (see below) every 200 ms (i.e., at a rate of 5 Hz; Figure 1A). The cloud's location mean was temporally independently resampled from five possible locations (-10°, -5°, 0°, 5°, 10°) on each trial with the inter-trial asynchrony jittered between 1.4 and 2.8 s in steps of 200 ms. In synchrony with the change in the cloud's location, the dots changed their colour and a concurrent sound was presented. The location of the sound was sampled from ± 5° visual angle with respect to the mean of the visual cloud. Observers' visual uncertainty estimate was quantified in terms of the relative weight of the auditory signal on the perceived sound location. The change in the dot's colour and the emission of the sound occurred in synchrony to enhance audiovisual binding.

Continuous sinusoidal and random walk sequences

Critically, to manipulate visual noise over time, the cloud's standard deviation changed at a rate of 5Hz according to i. a sinusoidal sequence, ii. a random walk sequence 1 or iii. a random walk sequence 2 (Figure 2). In all sequences the horizontal standard deviation of the visual cloud spanned a range from 2-18°:

i. *Experiment1 - Sinusoidal sequence (Sinusoid):* A sinusoidal sequence was generated with a period of 30s. During the ~65 min of the experiment, each participant completed ~ 130 cycles of the sinusoidal sequence.

ii. *Experiment2 - Random walk sequence 1 (RW1):* First, we generated a random walk sequence of 60 s duration using a Markov chain with 76 discrete states and transition probabilities of stay (1/3), change to lower (1/3) or upper (1/3) adjacent states. To ensure that the random walk sequence segment starts and ends with the same value, this initial 60 s sequence segment was concatenated with its temporally reversed segment resulting in a RW sequence segment of 120 s duration. Each participant was presented with this 120s RW1 sequence approximately 32 times during the experiment.

iii. *Experiment3 - Random walk sequence 2 (RW2):* Likewise, we created a second random-walk sequence of 15 s duration using a Markov chain with only 38 possible states and transition probabilities similar to above. The 15 s sequence was concatenated with its temporally reversed version resulting in a 30 s sequence. The smoothness of this sequence segment was increased by filtering it (without phase shift) with a moving average of 250 ms. Each participant was presented with this sequence segment ~130 times.

Generally, a session of a Sinusoid, RW1 or RW2 sequence included 1676 trials. Because of experimental problems, four sessions included only 1128, 1143 or 1295 trials. Before the experimental trials, participants practiced the auditory localization task in 25 unimodal auditory trials, 25 audiovisual congruent trials with a single dot as visual spatial cue and 75 trials with stimuli as in the main experiment.

Experiment 4 - Sinusoidal sequence with intermittent changes in visual noise (sinusoidal jump sequence)

To dissociate the Bayesian learner from approximate exponential discounting, we designed a sinusoidal sequence (period = 30 s) with intermittent increases / decreases in visual variance (Figure 5). As shown in Figure 5A, we inserted increases by 8° in visual STD at three levels of visual STD: 7.2°, 8.6°, 9.6° STD. Conversely, we inserted decreases by 8° in visual STD at

15.3°, 16.7°, 17.7° STD. We inserted jumps selectively in the period sections of high visual variance to make the jumps less apparent and maximize the chances that observers treated the series as a continuous sequence. As a result, the up-jumps occurred when the increases in visual variance were fastest (i.e. steeper slope), while the down-jumps occurred after sections in which the visual variance was relatively constant (i.e. shallow slope). We factorially combined these 3 (increases) x 3 (decreases) such that each sinewave cycle included exactly one sudden increase and decrease in visual STD (i.e., 9 jump types). Otherwise, the experimental paradigm and stimuli were identical to the continuous sinusoidal sequence described above. During the ~80 min of this experiment, each participant completed ~ 154 cycles of the sinusoidal sequence including 16-18 cycles for each of the 9 jump types. This sinusoidal jump sequence was expected to maximize differences in adaptation rate for the Bayesian and exponential learner. If participants continuously update their estimates of the visual reliability, as opposed to using a change point model (Adams and MacKay, 2007, Heilbron and Meyniel, 2019), the exponential learner will weight past and present uncertainty estimates throughout the entire sequence according to the same exponential function. By contrast, the Bayesian learner will take into account its uncertainty about the visual reliability and therefore adapt its visual reliability estimate for jumps from high to low visual variance (resp. low to high visual reliability, see Figure 6) more slowly than the exponential learner (see Appendix 1).

Subject numbers and inclusion criteria

30 of the 76 subjects participated in the sinusoidal and the RW1 sequence session. Eight additional subjects participated only in the RW1 sequence session. 18 additional subjects participated in the RW2 sequence session. One participant completed all three continuous sequences. 20 subjects participated in the sinusoidal sequence with intermittent changes in visual uncertainty. In total, we collected data from 30 participants for the sinusoidal, 38 participants for the RW1, 19 participants for the RW2 and 20 participants for the sinusoidal jump sequence. The sample sizes of 20-38 participants were based on a pilot experiment, which showed individually significant effects of past visual noise on the weighting of audiovisual spatial signals in 6/6 pilot participants. From these samples, we excluded participants if their perceived sound location did not depend on the current visual reliability (i.e., inclusion criterion $p < 0.05$ in the linear regression; please note that this inclusion criterion is orthogonal to the question of whether participants' visual uncertainty estimate depends on visual signals prior to the current trial). Thus, we excluded five participants of the sinusoidal and RW1 sequence and two participants from the sinusoidal jump sequence. Finally, we analysed data from 25

participants for the sinusoidal, 33 participants for the RW1, 19 participants for the RW2 and 18 participants for the sinusoidal jump sequence.

Experimental setup

Audiovisual stimuli were presented using Psychtoolbox 3.09 (Brainard, 1997, Kleiner et al., 2007) (www.psychtoolbox.org) running under Matlab R2010b (MathWorks) on a Windows machine (Microsoft XP 2002 SP2). Auditory stimuli were presented at ~75 dB SPL using headphones (Sennheiser HD 555). As visual stimuli required a large field of view, they were presented on a 30" LCD display (Dell UltraSharp 3007WFP). Participants were seated at a desk in front of the screen in a darkened booth, resting their head on an adjustable chin rest. The viewing distance was 27.5 cm. This setup resulted in a visual field of approx. 100°. Participants responded via a standard QWERTY keyboard. Participants used the buttons [i, 9, 0, -, =] with their right hand for localization responses.

Data analysis

Continuous sinusoidal and random walk sequences

At trial onset the visual cloud's location mean was independently resampled from five possible locations (-10°, -5°, 0°, 5°, 10°). Concurrently, the cloud's dots changed their colour and a sound was presented sampled from ± 5° visual angle with respect to the mean of the visual cloud. The inter-trial asynchrony was jittered between 1.4 and 2.8 s in steps of 200 ms. Therefore, across the experiment the trial onsets occurred at different times relative to the period of the changing visual cloud's STD resulting in a greater effective sampling rate than provided if the inter-trial asynchrony had been fixed.

For each period of the three continuous sinusoidal and random walk sequences, we sorted the trials (i.e. trial-specific visual cloud's STD, visual location, auditory location and observers' sound localization responses) into 20 temporally adjacent bins that covered one complete period of the changing visual STD. This resulted in about 1676 trials in total/20 bins = approximately 80 trials on average per bin in each subject (more specifically: a range of 52-96 (Sin), 52-92 (RW 1) or 71-93 (RW2) trials, for details see Supplementary file 1-Table 1).

We quantified the influence of the auditory and visual locations on observers' perceived auditory location for each bin by estimating a regression model separately for each bin (i.e. one regression model per bin). For instance, for bin = 1 we computed:

(1) $$R_{A,trial,bin=1} = L_{A,trial,bin=1}\, ß_{A,bin=1} + L_{V,trial,bin=1}\, ß_{V,bin=1} + ß_{const,bin=1} + e_{trial,bin=1}$$

with $R_{A,trial,bin=1}$ = Localization response for trial t and bin 1; $L_{A,trial,bin=1}$ or $L_{V,trial,bin=1}$ = 'true' auditory or visual location for trial t and bin 1; $ß_{A,bin=1}$ or $ß_{V,bin=1}$ = auditory or visual weight for bin 1; $ß_{const,bin=1}$ = constant term; $e_{trial,bin=1}$ = error term for trial t and bin 1. For each bin b, we thus obtained one auditory and one visual weight estimate. The *relative* auditory weight for a particular bin was computed as $w_{A,bin} = ß_{A,bin} / (ß_{A,bin} + ß_{V,bin})$ (Figure 2A-C).

By design, the temporal evolution of the physical visual variance (i.e., STD of the visual cloud) is symmetric for each period in the sinusoidal, RW1 and RW2 sequences. In other words, for physical visual noise, the 1st half and the flipped 2nd half within a period are identical (Figure 3E). Given this symmetry constraint, we evaluated the influence of past visual noise on participants' auditory weight $w_{A,bin}$ by comparing the w_A for the bins in the 1st half and the flipped 2nd half in a repeated measures ANOVA. If human observers estimate visual uncertainty by combining prior with current visual uncertainty estimates as expected for a Bayesian learner, w_A should differ between the 1st half and the mirror-symmetric flipped 2nd half of the sequence. More specifically, w_A should be smaller for the 1st half in which visual variance increased than for the mirror-symmetric time points of the 2nd half in which visual variance decreased. To test this prediction, we entered the subject-specific $w_{A,bin}$ into 2 (1st vs. flipped 2nd half) x 9 (bins, i.e. removing the bins at maximal and minimal visual noise values) repeated measures ANOVAs separately for the sinusoidal, RW1 and RW2 experiments (Table 1). For the sinusoidal sequence, we expected a main effect of 'half' because the sequence increased/decreased monotonically within each half period. For the RW1 and RW2 sequences, an influence of prior visual noise might also be reflected in an interaction effect of 'half x bin' because these sequences increased/decreased non-monotonically within each half period.

To further test whether the noise of past visual signals influenced observers' current visual uncertainty estimate, we employed a regression model in which the relative auditory weights $w_{A,bin}$ were predicted by the visual STD in the current bin and the difference in STD between the current and the previous bin:

(2) $$w_{A,bin} = \sigma_{V,bin}\ ß_{\sigma V} + (\sigma_{V,bin} - \sigma_{V,bin-1})\ ß_{\Delta\sigma V} + ß_{const} + e_{bin}$$

with $w_{A,bin}$= relative auditory weight in bin b; $\sigma_{V,bin}$ = mean visual STD in current bin b or previous bin b-1; $ß_{const}$ = constant term; e_{bin} = error term. To allow for generalization to the population level, the parameter estimates ($ß_{\sigma V}$, $ß_{\Delta\sigma V}$) for each participant were entered into two-sided one-sample t-tests at the between-subject random-effects level.

Sinusoidal sequence with intermittent changes in visual uncertainty

For each period of the sinusoidal sequence with intermittent changes, we sorted the values for the physical visual cloud's variance (i.e., the cloud's STD) and sound localization responses into 15 temporally adjacent bins which were positioned to capture the jumps in visual noise. For analysis of these sequences, we recombined the first and second halves of the 3 (increases at low, middle, high) x 3 (decreases at low, middle, high) sinewave cycles into three types of sinewave cycles such that both jumps were at low (= outer jump), middle (=middle jump) or high (= inner jump) visual noise. This recombination makes the simplifying assumption that the jump position of the first half will have negligible effects on participants' uncertainty estimates of the second half. As a result of this recombination, each bin comprised at least 44-51 trials across participants (Supplementary file 1-Table 1). As for the continuous sequences, we quantified the auditory and visual influence on the perceived auditory location for each bin based on separate regression models for the 15 temporally adjacent bins. For instance, for bin = 1 we computed: $R_{A,trial,\,bin=1} = L_{A,trial,bin=1} * ß_{A,bin=1} + L_{V,trial,bin=1} * ß_{V,bin=1} + ß_{const,bin=1} + e_{trial,bin=1}$. Next, we independently computed the relative auditory weight $w_{A,bin} = ß_{A,bin} / (ß_{A,bin} + ß_{V,bin})$ for each of the 15 temporally adjacent bins. We statistically evaluated the influence of past visual noise on participants' auditory weight on the w_A in terms of the difference between 1st half and flipped 2nd half using a 2 (1st vs. flipped 2nd half) x 7 (bins) x 3 (jump: inner, middle, outer) repeated measures ANOVAs (Table 1).

Computational Models (for continuous and discontinuous sequences)

To further characterize whether and how human observers use their uncertainty about previous visual signals to estimate their uncertainty of the current visual signal, we defined and compared three models in which visual reliability (λ_V) was (1) estimated instantaneously for each trial (i.e., instantaneous learner), was updated via (2) Bayesian learning or (3) exponential discounting (i.e. exponential learner) (Figure 1-figure supplement 1).

In the following, we will first describe the generative model that accounts for the fact that (1) visual uncertainty usually changes slowly across trials (i.e. time-dependent uncertainty changes) and (2) auditory and visual signals can be generated by one common or two independent sources (i.e. causal structure). Using this generative model as a departure point, we then describe how the instantaneous learner, the Bayesian learner and the exponential learner perform inference. Finally, we will explain how we account for participants' internal noise and predict participants' responses from each model (i.e. the experimenter's uncertainty).

Generative model

On each trial *t*, the subject is presented with an auditory signal A_t, from a source $S_{A,t}$, (see Figure 1-figure supplement 1) together with a visual cloud of dots at time *t* arising from a source, $S_{V,t}$, drawn from a Normal distribution $S_{V,t} \sim N(0, 1/\lambda_S)$ with the spatial reliability (i.e., inverse of the spatial variance): $\lambda_S = 1/\sigma_S{}^2$. Critically, $S_{A,t}$ and $S_{V,t}$, can either be two independent sources (C = 2) or one common source (C=1): $S_{A,t} = S_{V,t} = S_t$ (Kording et al., 2007).

We assume that the auditory signal is corrupted by noise, so that the internal signal is $A_t \sim N(S_{A,t}, 1/\lambda_A)$. By contrast, the individual visual dots (presented at high visual contrast) are assumed to be uncorrupted by noise, but presented dispersed around the location $S_{V,t}$ according to $V_{i,t} \sim N(U_t, 1/\lambda_{V,t})$, where $U_t \sim N(S_{V,t}, 1/\lambda_{V,t})$. The dispersion of the individual dots, $1/\lambda_{V,t}$, is assumed to be identical to the uncertainty about the visual mean, allowing subjects to use the dispersion as an estimate of the uncertainty about the visual mean.

The visual reliability of the visual cloud, $\lambda_{V,t} = 1/\sigma_{V,t}{}^2$, varies slowly at the re-display rate of 5 Hz according to a log random walk: $\log \lambda_{V,t} \sim N(\log \lambda_{V,t-1}, 1/\kappa)$ with $1/\kappa$ being the variability of $\lambda_{V,t}$ in log space. We also use this log random walk model to approximate learning in the four jump sequence (see (Behrens et al., 2007)).

The generative models of the instantaneous, Bayesian and exponential learners all account for the causal uncertainty by explicitly modeling the two potential causal structures. Yet, they differ in how they estimate the visual uncertainty on each trial, which we will describe in greater detail below.

Observer Inference

The instantaneous, Bayesian and exponential learners invert this (or slightly modified, see below) generative model during perceptual inference to compute the posterior probability of the auditory location, $S_{A,t}$, given the observed A_t and $V_{i,t}$. The observer selects a response based on the posterior using a subjective utility function which we assume to be the minimization of the squared error $(S_{A,t} - S_{true})^2$. For all models, the estimate for the location of the auditory source is obtained by averaging the auditory estimates under the assumption of common and independent sources by their respective posterior probabilities (i.e. model averaging, see Figure 1-figure supplement 1):

(3) $$\hat{S}_{A,t} = \hat{S}_{A,C=1,t}\, P(C_t = 1|A_t, V_{1:n,t}) + \hat{S}_{A,C=2,t}(1 - P(C_t = 1|A_t, V_{1:n,t}))$$

where $\hat{S}_{A,C=1,t}$ and $\hat{S}_{A,C=2,t}$ depend on the model (see below), and $P(C = 1|A_t, V_{1:n,t})$ is the posterior probability that the audio and visual stimuli originated from the same source according to Bayesian causal inference (Kording et al., 2007).

(4) $P(C_t = 1|A_t, V_{1:n,t}) =$

$$\frac{(P(A_t, V_{1:n,t}|C = 1)P(C_t = 1))}{\left(P(A_t, V_{1:n,t}\middle|C_t = 1)P(C_t = 1) + \left(P(A_t, V_{1:n,t}\middle|C_t = 2\right)(1 - P(C_t = 1))\right)}$$

Finally, for all models we assume that the observer pushes the button associated with the position closest to $\hat{S}_{A,t}$. In the following, we describe the generative and inference models for the instantaneous, Bayesian and exponential learners. For the Bayesian learner, we focus selectively on the model component that assumes a common cause, C = 1 (for full derivation including both model components, see Appendix 2.

Model 1: Instantaneous learner

The instantaneous learning model ignores that the visual reliability (i.e., the inverse of visual uncertainty) of the current trial depends on the reliability of the previous trial. Instead, it estimates the visual reliability independently for each trial from the spread of the cloud of visual dots:

(5) $P(S_{A,t}, U_t, \lambda_{\mathrm{V,t}}|\, A_{1:t}, V_{1:n,1:t}) = P(S_{A,t}, U_t, \lambda_{\mathrm{V,t}}|\, A_t, V_{1:n,t}) =$

$P(C = 1|A_t, V_{1:n,t})P_{C=1}(S_t, U_t, \lambda_{\mathrm{V,t}}|\, A_t, V_{1:n,t}) +$

$P(C = 2|A_t, V_{1:n,t})P_{C=2}(S_{A,t}, U_t, \lambda_{\mathrm{V,t}}|\, A_t, V_{1:n,t}) =$

$$P(C = 1|A_t, V_{1:n,t})\frac{P(S_t)P(A_t|S_t)P_{C=1}(U_t|S_t, \lambda_{\mathrm{V,t}})P(V_{1:n,t}|U_t, \lambda_{\mathrm{V,t}})P(\lambda_{\mathrm{V,t}})}{Z_1} +$$

$\left(1 - P(C = 1|A_t, V_{1:n,t})\right) P(S_{A,t})P(A_t|S_{A,t})/Z_2.$

with Z_1, Z_2 as normalization constants.

Apart from $P(C = 1|A_t, V_t)$, these terms are all normal distributions, while we assume in this model that $P(\lambda_{\mathrm{V,t}})$ is uninformative. Hence, visual reliability is computed from the variance: $\widehat{\lambda_{Vt}} = 1/(\sigma_{Vt}^2 + \frac{\sigma_{Vt}^2}{n})$ where $\sigma_{Vt}^2 = 1/(n-1)\sum_{i=1}^{n}(V_{i,t} - \bar{V}_{i,t})^2$ is the sample variance (and $\bar{V}_t = 1/n\sum_{i=1}^{n} V_{i,t}$ is the sample mean). The causal component estimates are given by:

(6) $$\hat{S}_{A,C=1,t} = \frac{\hat{\lambda}_{V,t}\bar{V}_t + \lambda_A \mathrm{A}_t}{\lambda_{V,t}+\lambda_A+\lambda_S}$$

(7) $$\hat{S}_{A,C=2,t} = \frac{\lambda_A \mathrm{A}_t}{\lambda_A+\lambda_S}$$

These two components are then combined based on the posterior probabilities of common and independent cause models (see equation 3). This model is functionally equivalent to a Bayesian causal inference model as described in Koerding et al. (2007), but with visual reliability

computed directly from the sample variance rather than a fixed unknown parameter (which the experimenter estimates during model fitting).

Model 2: Bayesian learner

The Bayesian learner capitalizes on the slow changes in visual reliability across trials and combines past and current inputs to provide a more reliable estimate of visual reliability and hence auditory location. It computes the posterior probability based on all auditory and visual signals presented until time t (here only shown for C=1, see Appendix 2.

According to Bayes rule, the joint probability of all variables until time *t* can be written based on the generative model as:

(8) $$P\left(\lambda_{\mathrm{V},1:\mathrm{t}}, A_{1:t}, U_{1:t}, V_{1:n,1:t}, S_{1:t}\right) =$$

$$P(A_1|S_1)P\left(V_{1:n,1}\middle|U_1, \lambda_{\mathrm{V},1}\right)P\left(U_1\middle|S_1, \lambda_{\mathrm{V},1}\right)P(S_1)P\left(\lambda_{\mathrm{V},1}\right)$$

$$\prod_{k=2}^{t} P(A_k|S_k)P\left(V_{1:n,k}\middle|U_k, \lambda_{\mathrm{V,k}}\right)P\left(U_k\middle|S_k, \lambda_{\mathrm{V,k}}\right)P\left(\lambda_{\mathrm{V,k}}|\lambda_{\mathrm{V,k-1}}\right)P(S_k)$$

As above, the visual likelihood is given by the product of individual Normal distributions for each dot *i*: $P\left(V_{1:n,t}\middle|U_t, \lambda_{\mathrm{V,t}}\right) = \prod_{i=1}^{n} N(V_{i,t}|U_t, 1/\lambda_{\mathrm{V,t}})$, and $P\left(U_t\middle|S_t, \lambda_{\mathrm{V,t}}\right) = N(U_t|S_t, 1/\lambda_{\mathrm{V,t}})$. The prior $P(S_t)$ is a Normal distribution $N(S_t|0, 1/\lambda_{\mathrm{S}})$ and the auditory likelihood $P(A_{t,}|S_t)$ is a Normal distribution $N(A_t|S_t, 1/\lambda_A)$. As described in the generative model, $P\left(\lambda_{\mathrm{V,k}}|\lambda_{\mathrm{V,k-1}}\right)$ is given by $\log\lambda_{\mathrm{V,t}} \sim N(\log\lambda_{\mathrm{V,t-1}}, 1/\kappa)$.

Importantly, only the visual reliability, $\lambda_{\mathrm{V,t}}$, is directly dependent on the previous trial ($P\left(\lambda_{\mathrm{V,k}}, \lambda_{\mathrm{V,k-1}}\right) = P\left(\lambda_{\mathrm{V,k}}|\lambda_{\mathrm{V,k-1}}\right)P\left(\lambda_{\mathrm{V,k-1}}\right) \neq P\left(\lambda_{\mathrm{V,k}}\right)P\left(\lambda_{\mathrm{V,k-1}}\right)$). Because of the Markov property (i.e. $\lambda_{\mathrm{V,t}}$ depends only on $\lambda_{\mathrm{V,t-1}}$), the joint distribution for time *t* can be written as

(9) $$P(\lambda_{\mathrm{V,t}}, \lambda_{\mathrm{V,t-1}}, A_t, U_t, V_{1:n,t}, S_t) =$$

$$P(A_t|S_t)P\left(U_t\middle|S_t, \lambda_{\mathrm{V,t}}\right)P\left(V_{1:n}\middle|U_t, \lambda_{\mathrm{V,t}}\right)P\left(\lambda_{\mathrm{V,t}}|\lambda_{\mathrm{V,t-1}}\right)P\left(\lambda_{\mathrm{V,t-1}}|V_{1:n,t-1}, A_{t-1}\right)P(S_t)$$

.

Hence, the joint posterior probability over location and visual reliability given a stream of auditory and visual inputs can be rewritten as:

(10) $$P(S_t, U_t, \lambda_{\mathrm{V,t}}|\, A_{1:t}, V_{1:n,1:t}) =$$

$$P(S_t)P(A_t|S_t)P\left(U_t\middle|S_t, \lambda_{\mathrm{V,t}}\right)P\left(V_{1:n}\middle|U_t, \lambda_{\mathrm{V,t}}\right)\int P\left(\lambda_{\mathrm{V,t}}|\lambda_{\mathrm{V,t-1}}\right)P\left(\lambda_{\mathrm{V,t-1}}|V_{1:n,t-1}, A_{t-1}\right)d\lambda_{\mathrm{V,t-1}}/$$

$$Z.$$

As this equation cannot be solved analytically, we obtain an approximate solution by factorizing the posterior in terms of the unknown variables ($S_t, U_t, \lambda_{\mathrm{V,t}}$) according to the method of variational Bayes (Bishop, 2006). In this approximate method (for details see Appendix 2), the

posterior is factorized into three terms, each a normal distribution: $P\left(S_t, U_t, \lambda_{\mathrm{V,t}_t} \middle| A_{\mathrm{t}}, V_{1:\mathrm{n,t}}\right) \approx q\left(S_t, U_t, \lambda_{\mathrm{V,t}_t}\right) = q(S_t) * q(U_t) * q\left(\lambda_{\mathrm{V,t}_t}\right)$.

In order to estimate the set of parameters (mean and variance) of $q(S_t)$, $q(U_t)$ and $q\left(\lambda_{\mathrm{V,t}_t}\right)$, the Free Energy is minimized iteratively (and thereby the Kullback–Leibler divergence between the true and approximate distribution), until a convergence criterion is reached (here, the change in each fitted parameter is less than 0.0001 between iterations).

This is done separately for the common cause model component (C=1) and the independent cause model component (C=2). The auditory location, for the common cause model is based on the approximation over the posterior location of $\hat{S}_{A,C=1,t}$ from, $q_1(S_t) = N(\hat{S}_{A,C=1,t}, \sigma_{1,t})$. The auditory location for the independent cause model is simply computed as $\hat{S}_{A,C=2,t} = A_t/(1 + \sigma_A^2/\sigma_0^2)$, because it is independent of the visual signal.

The marginal model evidence is estimated based on the minimized Free Energy for each mode component, $P(A_t, V_{1:n,t} | C = 1)$, respectively $P(A_t, V_{1:n,t} | C = 2)$ to form the posterior probability $P(C = 1 | A_t, V_{1:n,t})$, as described above in equation 4. These values can then be used to compute the predicted responses for a particular participant according to equation 3.

Model 3: Exponential learner

Finally, the observer may approximate the full Bayesian inference of the Bayesian learner by a more simple heuristic strategy of exponential discounting. In the exponential discounting model, the observer learns the visual reliability by exponentially discounting past visual reliability estimates:

(11) $\hat{\lambda}_{V,t-1} = 1/\sigma_{Vt}^2\,(1 - \gamma) + \hat{\lambda}_{V,t-1}\gamma$

where $\sigma_{Vt}^2 = 1/(\mathrm{n} - 1)\sum_{i=1}^{n}(V_{i,t} - \bar{V}_{i,t})^2$ is the sample variance and $\bar{V}_t = 1/\mathrm{n}\sum_{i=1}^{n} V_{i,t}$ is the sample mean.

Similar to the optimal Bayesian learner (above), this observer model uses the past to compute the current reliability, but it does so based on a fixed learning rate 1 - γ. Computation is otherwise performed in accordance with models 1 and 2, equations 3-4 and 6-7.

Assumptions of the computational models: motivation and caveats

Computational models inherently make simplifying assumptions about the generation of the sensory inputs and observers' inference.

First, we modelled that visual signals (i.e. the cloud's mean) were sampled from a Gaussian, while they were sampled from a uniform discrete distribution (i.e. [-10°, -5°, 0°, 5°, 10°]) in the experiment. Gaussian assumptions about the stimuli locations have nearly exclusively been made in the recent series of studies focusing on Bayesian Causal Inference in multisensory perception (Kording et al., 2007, Rohe and Noppeney, 2015b, Rohe and Noppeney, 2015a). Because visual signals have been sampled from a wide range of visual angle (i.e. 20°) and are corrupted by physical (i.e. cloud of dots) and internal neural noise, we used the simplifying assumption of a Gaussian spatial prior consistent with previous research.

Second, we assumed that the auditory signal location is sampled from a Gaussian, while the experiments presented sounds ± 5° from the visual location. These Gaussian assumptions about sound location can be justified by the fact that observers are known to be limited in their sound localization ability, particularly when generic head related transfer functions were used to generate spatial sounds. Moreover, because sounds are presented together with visual signals, it is even harder for observers to obtain an accurate estimate of the sound's location.

Third, in the experiment we generated the cloud of dots directly from a Gaussian distribution centred on S_t. By contrast, in the model we introduced a hidden variable U_t that is sampled from a Gaussian centred on S_t. The visual cloud of dots is then centred on this hidden variable U_t. We introduced this additional hidden variable U_t to account for observers' additional causal uncertainty in natural environments, in which even signals from a common source may not fully coincide in space. Critically, the dispersion of the cloud of dots is set to be equal to the standard deviation of the distribution from which U_t is sampled, so that the cloud's standard deviation informs observers about the variance of the hidden variable U_t.

Inference by the experimenter

From the observer's viewpoint, this completes the inference process. However, from the experimenter's viewpoint, the internal variable for the auditory stimulus, A_t, is unknown and not directly under the experimenter's control. To integrate out this unknown variable, we

generated 1,000 samples of the internal auditory value for each trial from the generative process $A_t \sim N(S_{A,t,true}, \sigma_A^2)$, where $S_{A,t,true}$ was the true location the auditory stimulus came from. For each value of A_t, we obtained a single estimate $\hat{S}_{A,t}$ (as described above). To link these estimates with observers' button response data, we assumed that subjects push the button associated with the position closest to $\hat{S}_{A,t}$. In this way, we obtained a histogram of responses for each subject and trial which provide the likelihood of the model parameters given a subject's responses: $P(resp_t|\kappa, \sigma_A, P_{common}, S_{A,t,true}, S_{V,t,true})$.

Model estimation and comparison

Parameters for each model (for all models: σ_A, P_{common}= P(C=1), σ_0, Bayesian learner: κ, exponential learner: γ) were fit for each individual subject by sampling using a symmetric proposal Metropolis-Hasting (MH) algorithm (with A_t integrated out via sampling, see above). The MH algorithm iteratively draws samples set_n from a probability distribution through a variant of rejection sampling: if the likelihood of the parameter set is larger than the previous set, the new set is accepted, otherwise it is accepted with probability $L(model|set_n)/L(model|set_{n-1})$, where $L(resp|set_n) = \prod_t P(resp_t|\kappa, \sigma_A, P_{common}, S_{A,t,true}, S_{V,t,true})$ (for Bayesian learner). We sampled 4000 steps from 4 sampling chains with thinning (only using every 4th sampling to avoid correlations in samples), giving a total of 4000 samples per subject data sets. Convergence was assessed through scale reduction (using criterion R<1.1(Gelman et al., 2013)). Using sampling does not just provide a single parameter estimate for a data set (as when fitting maximum likelihood), but can instead be used to assess the uncertainty in estimation for the data set. The model code was implemented in Matlab (Mathworks, MA) and ran on two dual Xeon workstations. Each sample step, per subject data set, took 30 seconds on a single core (~42 hours per sampling chain).

Quantitative Bayesian model comparison of the three candidate models was based on the Watanabe-Akaike Information Criterion (WAIC) as an approximation to the out of sample expectation (Gelman et al., 2013). At the fixed-effects level, Bayesian model comparison was performed by summing the WAIC over all participants within each experiment. For a random-effects analysis, we transformed the WAIC into log-likelihoods by dividing them by minus 2. We then computed the protected exceedance probability that one model is better than the other model beyond chance using hierarchical Bayesian model selection (Penny et al., 2010, Rigoux et al., 2014).

To qualitatively compare the localization responses given by the participants and the responses predicted by the instantaneous, Bayesian and exponential learner, we computed the

auditory weight w_A from the predicted responses of the three models exactly as in the analysis for the behavioral data. For illustration, we show and compare the model's w_A from the 1st and the flipped 2nd half of the periods for each of the four experiments (cf. Figure 3, Figure 4, Figure 5B/C and Figure 5-figure supplement 1).

Parameter recovery

To test the validity of the models, we performed parameter recovery and were able to recover the generating values with a bias of all parameters smaller than 10 percent (for full details of bias and variance across parameters, see Appendix 1 and Supplementary file 1-Table 7).

Simulated localization responses

To further compare the Bayesian and exponential learner and assess whether they can be discriminated experimentally, we simulated the choices of 12 subjects for the continuous sinusoidal and sinusoidal jump sequence using the Bayesian learner model (parameters: σ_A =6 deg, κ=15, P_{common} =0.7 and σ_0 =12 degrees). To increase observers' uncertainty about their visual reliability estimates, we reduced the number of dots in the visual clouds from 20 to 5 dots where we ensured that the mean and variance of the 5 dots corresponded to the experimentally defined visual mean and variance. We then fitted the Bayesian learner and exponential learner models to each simulated data set (using the BADS toolbox for likelihood maximization (Acerbi and Ma, 2017)). The fitted parameters for the Bayesian model, set_{Bayes} were very close to the parameters used to generate observers' simulated responses (sinusoidal sequence, fitted parameters: σ_A = 6.11°, κ = 17.5, P_{common} = 0.72 and σ_0 = 12.4°; sinusoidal jump sequence, fitted parameters: σ_A = 6.08°, κ = 17.3 , P_{common} = 0.71 and σ_0 = 12.2°) – thereby providing a simple version of parameter recovery. The parameters of the exponential model, set_{Exp} (fitted to observers' responses generated from the Bayesian model) were very similar to those of the Bayesian learner (sinusoidal sequence: σ_A = 5.99°, γ =0.70, P_{common} =0.61 and σ_0 =12.0°, sinusoidal jump sequence: σ_A =6.06°, γ= 0.70 , P_{common} =0.65 and σ_0 =12.0°). Moreover, the fits to the simulated observers' responses were very close for the two models (Figure 6), with mean log likelihood difference ($log(L(resp|set_{Bayes}))$ – $\log(L(resp|set_{Exp}))$) =1.82 for the sinusoidal and 2.74 for the sinusoidal jump sequence (implying a slightly better fit for the Bayesian learner). Figure 6C and D show the timecourses of observers' visual uncertainty (STD) as estimated by the Bayesian and exponential learners.

Data availability

The behavioral data and model predictions as well as the code for modelling and analyses scripts are available in an OSF repository: https://osf.io/gt4jb/

References

ACERBI, L., DOKKA, K., ANGELAKI, D. E. & MA, W. J. 2018. Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLoS computational biology,* 14**,** e1006110.

ACERBI, L. & MA, W. J. 2017. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Advances in neural information processing systems (NIPS)***,** 1836-1846.

ACERBI, L., VIJAYAKUMAR, S. & WOLPERT, D. M. 2014. On the origins of suboptimality in human probabilistic inference. *PLoS computational biology,* 10**,** e1003661.

ADAMS, R. P. & MACKAY, D. J. 2007. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

ALAIS, D. & BURR, D. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol,* 14**,** 257-62.

ALGAZI, V. R., DUDA, R. O., THOMPSON, D. M. & AVENDANO, C. The cipic hrtf database. Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, 2001. IEEE, 99-102.

ALLER, M. & NOPPENEY, U. 2019. To integrate or not to integrate: Temporal dynamics of Bayesian Causal Inference. *PLoS Biol***,** in press.

BATTAGLIA, P. W., JACOBS, R. A. & ASLIN, R. N. 2003. Bayesian integration of visual and auditory signals for spatial localization. *J Opt Soc Am A Opt Image Sci Vis,* 20**,** 1391-7.

BECK, J. M., MA, W. J., KIANI, R., HANKS, T., CHURCHLAND, A. K., ROITMAN, J., SHADLEN, M. N., LATHAM, P. E. & POUGET, A. 2008. Probabilistic population codes for Bayesian decision making. *Neuron,* 60**,** 1142-1152.

BEHRENS, T. E., WOOLRICH, M. W., WALTON, M. E. & RUSHWORTH, M. F. 2007. Learning the value of information in an uncertain world. *Nature neuroscience,* 10**,** 1214.

BERNIKER, M., VOSS, M. & KORDING, K. 2010. Learning priors for Bayesian computations in the nervous system. *PLoS One,* 5.

BISHOP, C. M. 2006. *Pattern recognition and machine learning*, New York, Springer.

BRAINARD, D. H. 1997. The psychophysics toolbox. *Spatial vision,* 10**,** 433-436.

DRUGOWITSCH, J., DEANGELIS, G. C., KLIER, E. M., ANGELAKI, D. E. & POUGET, A. 2014. Optimal multisensory decision-making in a reaction-time task. *Elife,* 3**,** e03005.

DRUGOWITSCH, J., WYART, V., DEVAUCHELLE, A.-D. & KOECHLIN, E. 2016. Computational precision of mental inference as critical source of human choice suboptimality. *Neuron,* 92**,** 1398-1411.

ERNST, M. O. & BANKS, M. S. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature,* 415**,** 429-33.

FISER, J., BERKES, P., ORBAN, G. & LENGYEL, M. 2010. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn Sci,* 14**,** 119-30.

GELMAN, A., HWANG, J. & VEHTARI, A. 2014. Understanding predictive information criteria for Bayesian models. *Statistics and computing,* 24**,** 997-1016.

GELMAN, A., STERN, H. S., CARLIN, J. B., DUNSON, D. B., VEHTARI, A. & RUBIN, D. B. 2013. *Bayesian data analysis*, Chapman and Hall/CRC.

HEILBRON, M. & MEYNIEL, F. 2019. Confidence resets reveal hierarchical adaptive learning in humans. *PLoS computational biology,* 15**,** e1006972.

HOU, H., ZHENG, Q., ZHAO, Y., POUGET, A. & GU, Y. 2019. Neural Correlates of Optimal Multisensory Decision Making under Time-Varying Reliabilities with an Invariant Linear Probabilistic Population Code. *Neuron*.

JACOBS, R. A. 1999. Optimal integration of texture and motion cues to depth. *Vision Res,* 39**,** 3621-9.

JACOBS, R. A. & FINE, I. 1999. Experience-dependent integration of texture and motion cues to depth. *Vision Research,* 39**,** 4062-4075.

KLEINER, M., BRAINARD, D., PELLI, D., INGLING, A., MURRAY, R. & BROUSSARD, C. 2007. What's new in Psychtoolbox-3. *Perception,* 36**,** 1.1-16.

KNILL, D. C. & POUGET, A. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci,* 27**,** 712-9.

KNILL, D. C. & RICHARDS, W. 1996. *Perception as Bayesian inference*, Cambridge University Press.

KORDING, K. P., BEIERHOLM, U., MA, W. J., QUARTZ, S., TENENBAUM, J. B. & SHAMS, L. 2007. Causal inference in multisensory perception. *PLoS One,* 2**,** e943.
MA, W. J., BECK, J. M., LATHAM, P. E. & POUGET, A. 2006. Bayesian inference with probabilistic population codes. *Nat Neurosci,* 9**,** 1432-8.
MA, W. J. & JAZAYERI, M. 2014. Neural coding of uncertainty and probability. *Annual review of neuroscience,* 37**,** 205-220.
MEIJER, D., VESELIČ, S., CALAFIORE, C. & NOPPENEY, U. 2019. Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation. *Cortex,* 119**,** 74-88.
MIKULA, L., GAVEAU, V., PISELLA, L., KHAN, A. Z. & BLOHM, G. 2018. Learned rather than online relative weighting of visual-proprioceptive sensory cues. *Journal of neurophysiology,* 119**,** 1981-1992.
NORTON, E. H., ACERBI, L., MA, W. J. & LANDY, M. S. 2019. Human online adaptation to changes in prior probability. *PLoS computational biology,* 15**,** e1006681.
PENNY, W. D., STEPHAN, K. E., DAUNIZEAU, J., ROSA, M. J., FRISTON, K. J., SCHOFIELD, T. M. & LEFF, A. P. 2010. Comparing families of dynamic causal models. *PLoS Comput Biol,* 6**,** e1000709.
RIGOUX, L., STEPHAN, K. E., FRISTON, K. J. & DAUNIZEAU, J. 2014. Bayesian model selection for group studies—revisited. *Neuroimage,* 84**,** 971-985.
ROHE, T., EHLIS, A.-C. & NOPPENEY, U. 2019. The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. *Nature Communications,* 10**,** 1907.
ROHE, T. & NOPPENEY, U. 2015a. Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS Biol,* 13**,** e1002073.
ROHE, T. & NOPPENEY, U. 2015b. Sensory reliability shapes perceptual inference via two mechanisms. *Journal of Vision,* 15**,** 1-16.
ROHE, T. & NOPPENEY, U. 2016. Distinct computational principles govern multisensory integration in primary sensory and association cortices. *Current Biology,* 26**,** 509-514.
SHEN, S. & MA, W. J. 2016. A detailed comparison of optimality and simplicity in perceptual decision making. *Psychological review,* 123**,** 452.
TRIESCH, J., BALLARD, D. H. & JACOBS, R. A. 2002. Fast temporal dynamics of visual cue integration. *Perception,* 31**,** 421-434.
VAN BEERS, R. J., SITTIG, A. C. & GON, J. J. D. V. D. 1999. Integration of proprioceptive and visual position-information: An experimentally supported model. *Journal of neurophysiology,* 81**,** 1355-1364.
WOZNY, D. R., BEIERHOLM, U. R. & SHAMS, L. 2010. Probability matching as a computational strategy used in perception. *PLoS Comput Biol,* 6.
ZEMEL, R. S., DAYAN, P. & POUGET, A. 1998. Probabilistic interpretation of population codes. *Neural computation,* 10**,** 403-430.

Legends of supplementary figures

Figure 1-figure supplement 1. Generative model for the Bayesian learner. The Bayesian Causal Inference model explicitly models whether auditory and visual signals are generated by one common (C=1) or two independent sources (C=2) (for further details see Koerding et al., 2007). We extend this Bayesian Causal Inference model into a Bayesian learning model by making the visual reliability ($\lambda_{V,t}$, i.e. the inverse of uncertainty or variance) of the current trial dependent on the previous trial.

Figure 2-figure supplement 1. Time course of the relative auditory weights for continuous sequences of visual noise when controlling for location of the cloud of dots in the previous trial. Relative auditory weights (mean across participants ± SEM, left ordinate) and visual noise (i.e., STD of the cloud of dots, right ordinate) are displayed as a function of time as shown in Figure 2 of the main text. To compute the relative auditory weights, the sound localization responses where regressed on the A and V signal locations within bins of 1.5 s (**A**, **B**) or 6 s (**C**) width across sequence repetitions within each participant. To control for a potential effect of past visual locations, the location of the visual cloud of dots in the previous trial was included in this regression model as a covariate (cf. Supplementary file 1-Table 3).

Figure 5-figure supplement 1. Time course of relative auditory weights and visual noise for the sinusoidal sequence with intermittent jumps in visual noise. Relative auditory weights $w_{A,bin}$ (mean across participants) of the 1st (solid) and the flipped 2nd half (dashed) of a period (binned into 15 time bins) plotted as a function of the time in the sinusoidal sequence with intermitted inner (light gray), middle (gray) and outer (dark gray) jumps. Relative auditory weights were computed from auditory localization responses of exponential **(A)** or instantaneous **(B)** learning models. For comparison, the standard deviation of the visual signal is shown in **(C)**. Please note that all models were fitted to observers' auditory localization responses (i.e. not the auditory weight w_A).

Figure 5-figure supplement 2. Time course of relative auditory weights and root mean squared error of the computational models before and after the jumps in the sinusoidal sequence with intermittent jumps. (A) Relative auditory weights w_A (mean across participants) shown as a function of time around the up-jumps (left panel) and the down-jumps (right panel) for observers' behavior, the instantaneous, exponential and Bayesian learner. Relative auditory weights were computed from auditory localization responses for behavioral data and for the predictions of the three computational models in time bins of 200 ms (i.e., 5Hz rate of the visual clouds). Trials from the three types of up- and down-jumps were pooled to increase the reliability of the w_A estimates. Because time bins included only few trials in some participants, individual w_A values that were smaller or larger than the three times the scaled median absolute deviation were excluded from the analysis. Note that the up jumps occurred

around the steepest increase in visual noise, so that the Bayesian and exponential learners underestimated visual noise (cf. Figure 5C), leading to smaller w_A as compared to the instantaneous learner already before the up jump. (**B)** Root mean squared error (RMSE; computed across participants) between w_A computed from behavior and the models' predictions (as shown in A), shown as a function of the time around the up-jumps (left panel) and the down-jumps (right panel). Please note that all models were fitted to observers' auditory localization responses (i.e. not the auditory weight w_A).

Legends of appendix figures

Appendix 2-figure 1. Generative model, for one (C=1) or two sources (C=2).

Appendix 2-figure 2. Approximation of theta using Laplace approximation.

Appendix 2-figure 3. Compring variational Bayes approximation with a numerical discretised grid approximation. Top row: Example visual stimuli over eight subsequent trials. Middle row: The distribution of estimated sample variance, with no learning over trials. Bottom row: The distribution of _V;t for the Bayesian model that incorporates the learning across trials. Red line is the numerical comparison when using a discretised grid to estimate variance, as opposed to the variational Bayes (green line).

Acknowledgements

This study was funded by the ERC (ERC-multsens, 309349), the Max Planck Society and the Deutsche Forschungsgemeinschaft (DFG; grant number RO 5587/1-1).

Competing interests

The authors report no conflict of interest.